The Physical Habitat (PHAB) Index of Physical Integrity (IPI): Interim instructions for calculating scores using GIS and R

Andrew C. Rehn[1] (arehn@sbcglobal.net), Raphael Mazor[2] (raphaelm@scccwrp.org),
Peter R. Ode[1], Marcus Beck[2] (marcusb@sccwrp.org), Tyler Boyle[3], Erik Fintel[3],
Calvin Yang[4] (calvin.yang@waterboards.ca.gov)

[1]California Department of Fish and Wildlife. Rancho Cordova, CA

[2]Southern California Coastal Water Research Project. Costa Mesa, CA

[3]Geographical Information Center, California State University. Chico, CA

[4]State Water Resources Control Board. Sacramento, CA

SWAMP-SOP-2018-0006

Table of Contents

Introduction

This document describes steps in calculating the Physical Habitat (PHAB) Index of Physical Integrity (IPI), an assessment index that measures pysical habitat integrity based on PHAB metrics. The instructions provided herein are provided as interim support for analysts requiring IPI scores. The State Water Resources Control Board is currently developing a more automated approach to score calculation. Until that time, this document describes the only way to obtain IPI scores.

The first section in this document describes the process for using a geographic information system (GIS) to calculate environmental predictors, such as watershed area and rainfall. The second section describes the process for using the environmental predictors, as well as taxonomic data, to calculate IPI scores in R.

The development and interpretation of the index is described in SWAMP technical memo (https://www.waterboards.ca.gov/water_issues/programs/swamp/bioassessment/docs/physical_habitat_index_technical_memo.pdf):

> Rehn, A.C., R.D. Mazor and P.R. Ode. 2018. An Index to Measure the Quality of Physical Habitat in California Wadeable Streams. Swamp Technical Memorandum SWAMP-TM-2018-0005.

If you wish to cite this document to describe IPI calculation (as opposed to general index properties or development), use the following citation:

> A. C. Rehn, R. D. Mazor. P. R. Ode, M. Beck, T. Boyle, E. Fintel, and C. Yang. 2018. The Physical Habitat (PHAB) Index of Physical Integrity (IPI): Interim instructions for calculating scores using GIS and R. SWAMP-SOP-2018-0006.

Computer Software Requirements:

- ArcGIS 10.2.2 or higher
    - Spatial Analyst Extension (Extension for ArcGIS)
- NHDPlusV2 Basin Delineator V2.4.0.20 (instructions)
    - Microsoft .NET 4.6.10. or higher
    - Microsoft SQL Server 2012 Express LocalDB 64-bit
- R Studio 1.0.136 or higher (https://www.rstudio.com/products/rstudio/download/) or R 3.3.2 or higher (https://cran.r-project.org/)
- 7-Zip (https://www.7-zip.org/) file extaction software

Section 1: Instructions for Calculating PHAB Predictors with a Geographic Information System

The goal of this section is to guide users through the steps needed to calculate the predictors required for the Physical Habitat (PHAB) index.

These predictors are described as follows:

Predictor	Description
New_Lat	Latitude, in decimal degrees North
New_Long	Longitude, in decimal degrees East
SITE_ELEV	Site elevation in meters
ELEV_RANGE	Difference in elevation between the sample point and highest point in the catchment, in meters.
MAX_ELEV	Maximum elevation of the watershed, in meters
AREA_SQKM	Watershed area in square kilometers
PPT_00_09	Average precipitation (2000 to 2009) at the sample point, in hundredths of millimeters
KFCT_AVE	Average soil erodibility factor
MEANP_WS	Mean phosphorus geology from the watershed
MINP_WS	Minimum phosphorus geology from the watershed

Although the State Water Board will develop web-based tools to automate the steps described in this document, some users may be interested in calculating the PHAB on their own. We cannot guarantee the accuracy of metrics calculated using this document.

Field names and records are case-sensitive.

DOWNLOADING DATA

The necessary raster data can be downloaded from SWAMP's Google Drive.
The Watershed Metric Toolbox (top link) and the geodatabase (bottom link) can be downloaded using here:

https://drive.google.com/file/d/1U6Ns7sVgv5ihRLbwwf4zszMntUkbDA9G/view?usp=sharing
https://drive.google.com/file/d/19uUUG2dPhzCn93967uVcH2A-ptAfYevI/view?usp=sharing

This zip files contain a geodatabase, a python script for data consolidation and export, and documentation for each step in metric calculation.

CREATING THE BASEFILES

BaseFiles are shapefiles that function as the unit of spatial analysis for calculation of PHAB predictors and other spatial metrics. The PHAB predictors are calculated with two types of BaseFiles: The site (a point representing the sample location) and a catchment (a polygon representing the contributing landuse).

All BaseFiles must contain a unique identifier of each station, which we call "StationCod" (this field name gets automatically changed to "StationCode" when data are exported for analysis in R). StationCods must be represented in all shapefiles, using the same letter case, must not contain: periods, special characters, and spaces. Each StationCod must contain no more than 18 characters.

Creating the Site BaseFile

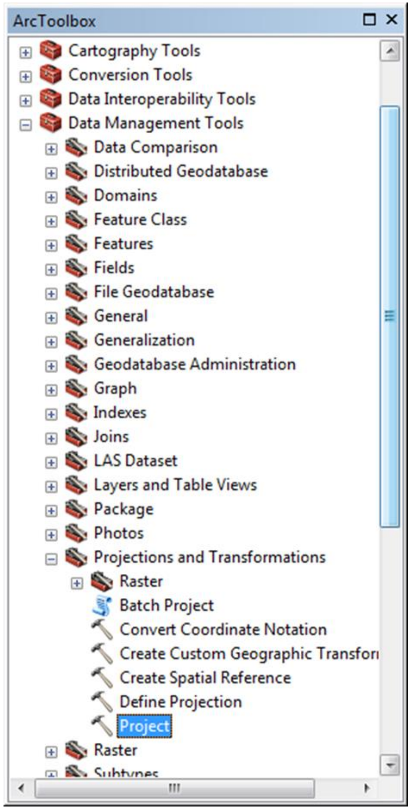
The goal of this step is to create a shapefile representing the location of sample points. Where possible, the location of sample points is adjusted ("snapped") from the actual coordinates to the nearest stream line represented in the National Hydrography Dataset Plus (NHD Plus). This "snapping" step is optional, but it is recommended because it improves the catchment delineation process, and also to help generate metrics for screening reference sites. If snapping is not desired, stop after Step 4, but be sure to give subsequent delineations, metrics, and other analytical products additional scrutiny.

Data requirements

-Spreadsheet (e.g., in .xls format) with unique site identifiers (field name: StationCode) and coordinates in decimal degrees (field names: LAT and LONG).

-NHDPlus V2 Data, including flowlines and Subwatersheds (Hydrologic Units). Full Data Requirements can be found in the NHDPlusV2 Basin Delineator readme. Data can be downloaded from the NHDPlusV2 database here. The flowline data are the file NHDPlusV21_CA_18_NHDSnapshot_05.7z and the Subwatersheds data are the file NHDPlusV21_CA_18_WBDSnapshot_03.7z. You will need 7-Zip software to extract the files.

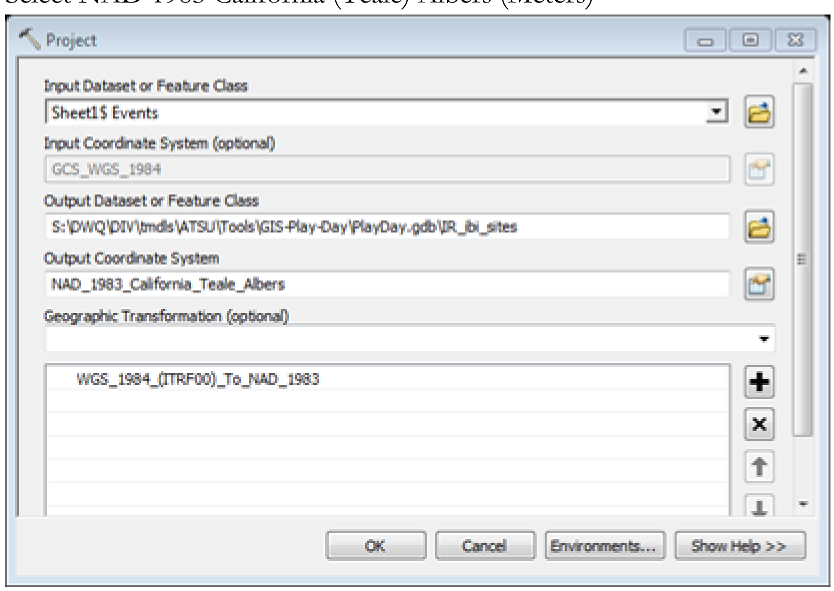
1. Load spreadsheet in ArcMap.
2. Right-click and display XY data. X field is the LONG, Y field is the LAT. Set the Coordinate system to WGS84.
    a. To get to WGS84:
        i. Expand Geographic Coordinate Systems
        ii. Expand World
        iii. Select WGS 1984
3. Reproject to NAD_1983_California_Teale_Albers.
    a. To reproject, open the ArcToolbox from the Geoprocessing menu or toolbar:

i. 

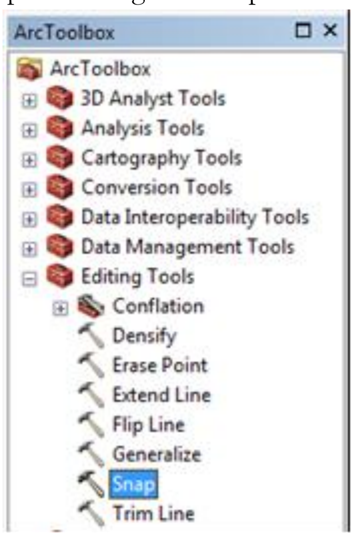
ii. Select point layer as "Input Dataset or Feature Class"

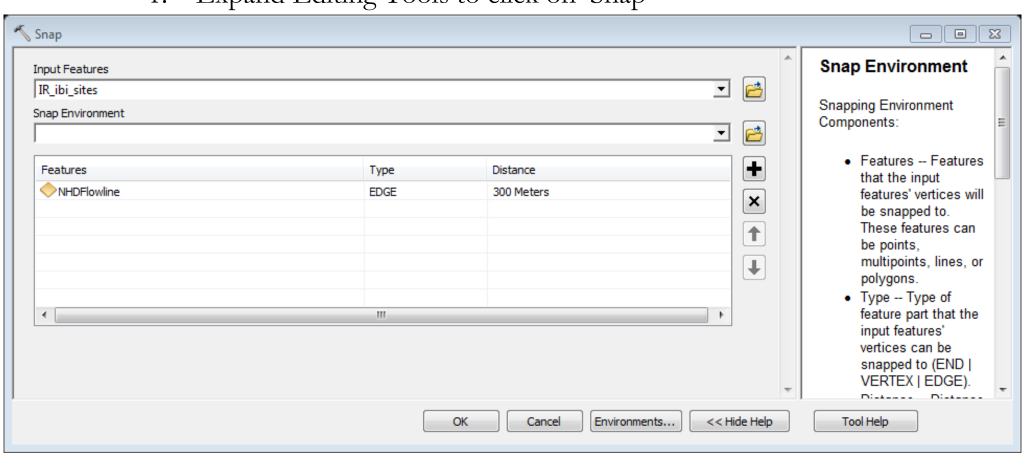
iii. Select Geodatabase you want to save to for "Output Dataset" field

iv. To fill out Output Coordinate System:

   1. Expand Projected Coordinate Systems
   2. Expand State Systems
   3. Select NAD 1983 California (Teale) Albers (Meters)
   4. 

Snapping points to NHD flowlines:

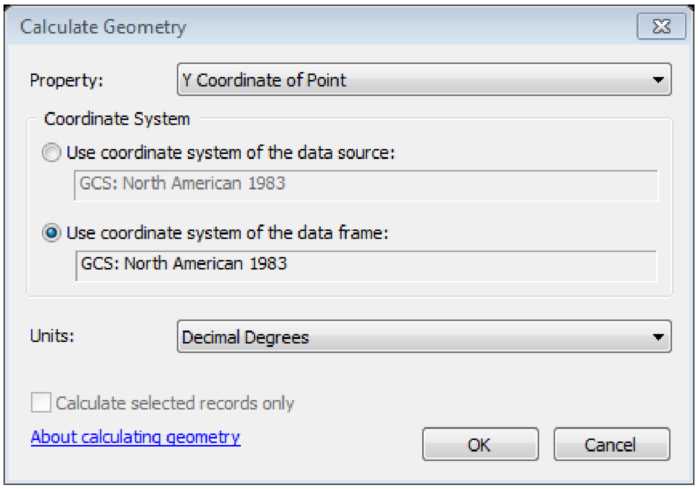
4. Load the spatially corresponding NHD flowlines and the Subbasin from the USGS website shown above. If all points fall within one NHD Region move on to the next steps. If not export them by region. They will need to be run separately through the delineator.
5. Snap the points to the nearest flowline in a manual edit session using "Edge Snapping". If there is no flowline near it, add a note in the attribute table. If the point is near a confluence or between two rivers look in the attribute table for clues to where it should go (see Quality Control section below).
6. To snap points using auto-snap tool from ArcToolbox:
    a. 

        i. To snap, open the ArcToolbox from the Geoprocessing menu or toolbar:
            1. Expand Editing Tools to click on 'Snap'
    b. 

        i. Select point layer as "Input Features"
        ii. Select NHDFlowline as "Snap Environment"
            1. Select EDGE as "Type"
            2. Enter 300 Meters as "Distance"
7. Once all sites are snapped, add a "New_Lat" and "New_Long" field. Calculate the latitude and longitude of the newly snapped points.
    a. To Add new Fields:
        i. Open Attribute Table
        ii. Select "Add Field…" under Table Options (top left icon in the table window)

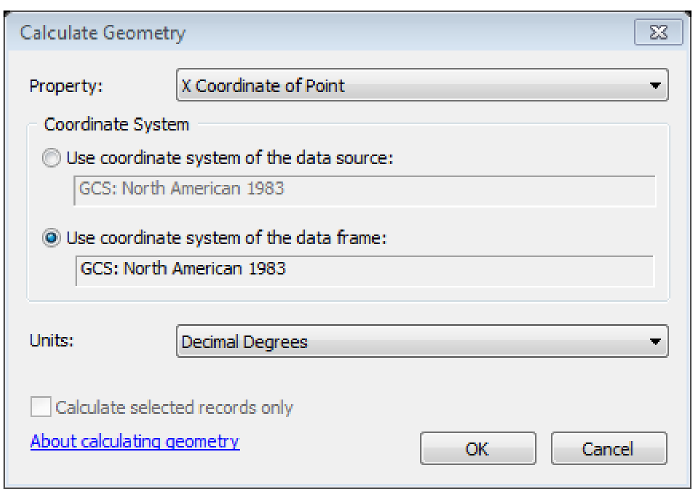
iii. Enter “New_Lat” for Name:
    1. Select “Double” under Type:
    2. Click OK
iv. Repeat for “New_Long”

b. To Calculate New_Lat:
    i. Right-click on New_Lat field name
    ii. Select Calculate Geometry…and click Yes on warning about Calculating outside of an edit session
    iii. Fill out menu as follows and click ok:



1. *Note: Latitude = Y coordinate

iv. To Calculate New_Long:
v. Right-click on New_Long field name
vi. Select Calculate Geometry…and click Yes on warning about Calculating outside of an edit session
vii. Fill out menu as follows:

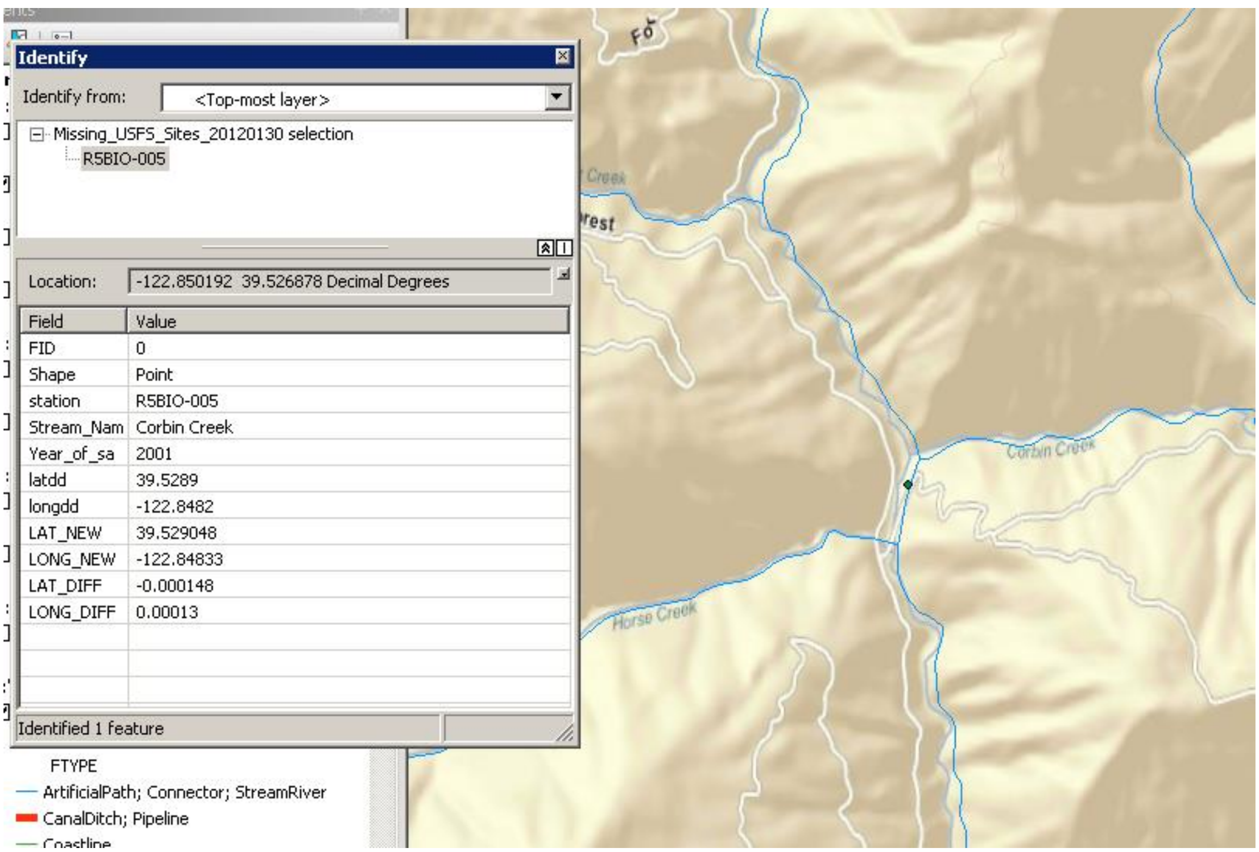


viii. Select OK and Yes on warning about calculating outside of an edit session
*Note: Longitude = X coordinate similar to what you did for Latitude.

8. Export the points and name them appropriately to make a permanent layer. Name this shapefile XXX_Sites (where XXX is the project name).

Quality control for the Sites Basefile (for both snapped and unsnapped sites).

1. Ensure that snapped locations are reasonably close to reported sampling locations (generally, less than 0.003 decimal degrees, or ~300 m on the ground). Sites that snapped large distances should be flagged, so that the catchments delineated later can receive additional review. Large snapping distances are not always problems and may have minimal impact on the catchment or the metrics calculated from the BaseFiles. In a few cases it can actually lead to an improvement in the position of a site (e.g., if the original coordinates plotted on a mountain side and the shift moved them down into the channel).
2. Look for ancillary data (such as station names or descriptions, aerial imagery, USGS topographic maps) to verify sampling location. Contact sampling crews if necessary.
3. For sites close to confluences or near transitional areas, close scrutiny is required to ensure that the site is located on the correct stream segment. In the figure below, a site was sampled on Corbin Creek, near the confluence with the Eel River (as indicated by the site name). However, the point plots on the main stem of the Eel, downstream of the confluence. The coordinates needs to be manually corrected.



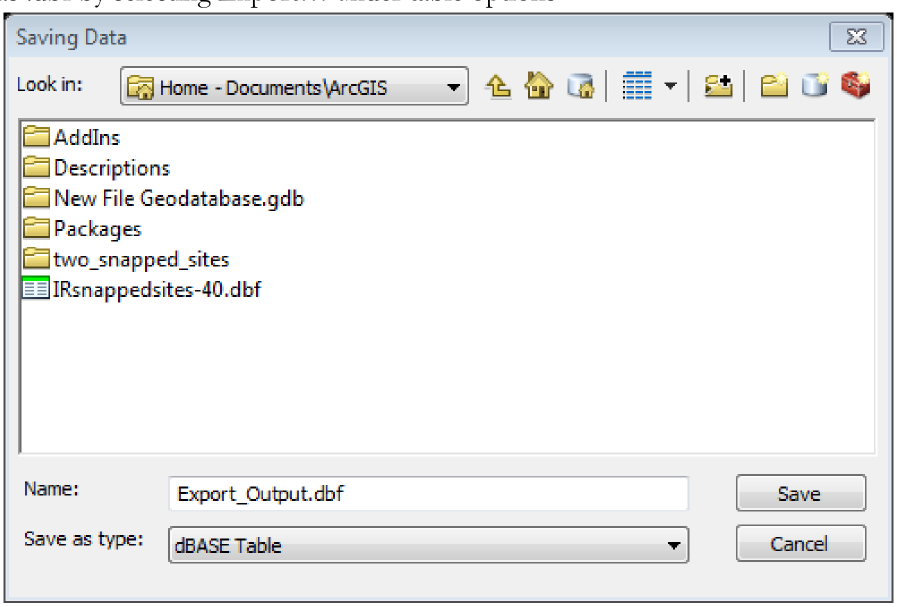
Creating the Catchments BaseFile

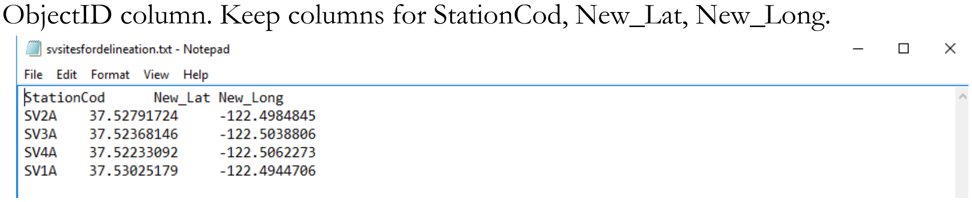
Below we outline the recommended approach for delineating catchments from a digital elevation model (DEM), simplified and improved by using pre-delineated watersheds in the National Hydrography Dataset Plus (NHD Plus). This approach works well for the majority of streams in California, although in certain situations alternative delineation methods may be preferable (particularly in flat areas with minimal topographic variation). No matter what approach is used, the goal is to identify the portion of the landscape that contributes runoff to a stream under natural ("reference") conditions. In general, dams, diversion, and inter-basin water transfers should be ignored when delineating the contributing catchment. This section describes three different ways to delineate catchments. The Basin Delineator is the preferred method followed by the ArcGIS Hydrology and Stream Stats methods. The Basin Delineator method is software that can be downloaded to user computers. The ArcGIS Hydrology method (snaps and delineates) is an ESRI online service that can only be accessed through an ArcGIS Online account. The Stream Stats method asks the user to submit sites online while returning zipped shapefiles back to the user.

<u>Basin Delineator:</u>

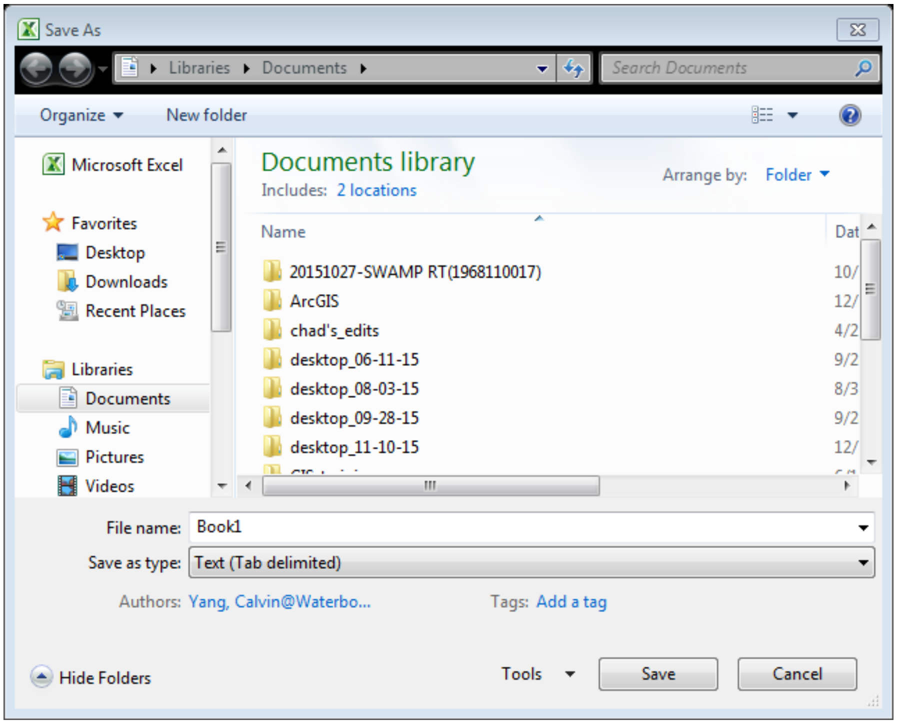
Requirements:

- Sites Base File
- 30-m DEM
- NHD Plus V2.1

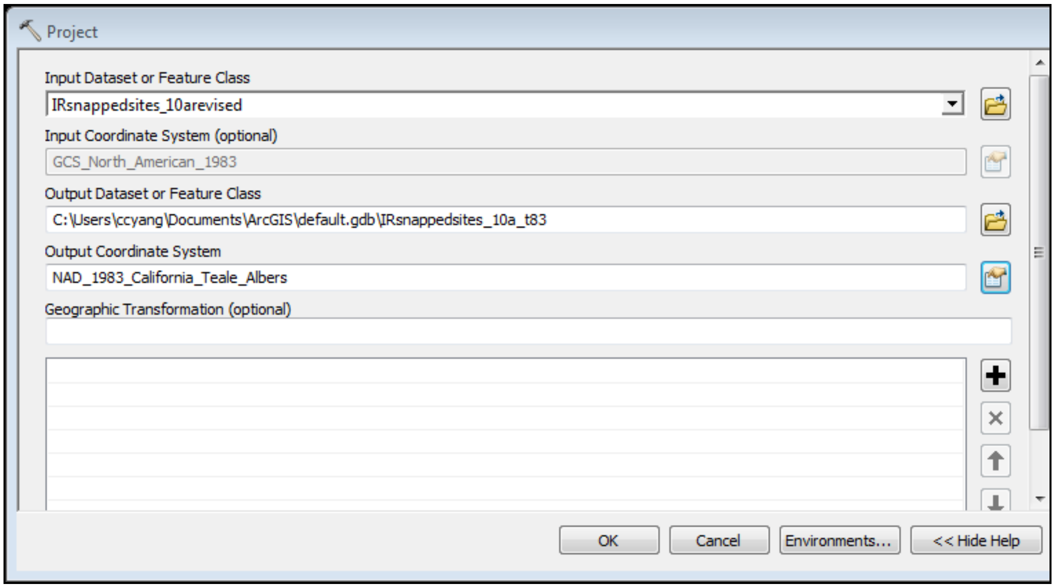
1. Load sites BaseFile.
2. Load the NHD flowlines and Subwatersheds from the Hydrologic Units folder from the appropriate region; watersheds in different regions must be delineated in separate batches.
3. Save the sites attribute table as a tab-delimited text file.
   a. Export as .dbf by selecting Export… under table options
      i. 

   b. Open .dbf file in Excel
   c. Delete ObjectID column. Keep columns for StationCod, New_Lat, New_Long.
      i. 

   d. Save the file as Text (tab delimited)

i. 

4. Copy the text file to the NHDPlus Tools working directory (local drive) on the processing computer
5. Start the basin delineator located here (version 24020): http://www.horizon-systems.com/nhdplus/NHDPlusV2_tools.php#NHDPlusV2 BasinDelineator Tool
    a. Before running the basin delineator, be sure that 'Split Catchments' is checked under System Setup so the delineation starts at the site rather than at the beginning of the catchment.
6. Click Run Basin Delineator.
7. The "Basin Pourpoints File" is the text file you made; browse to it.
8. Set the "Basin Shape Output File" to an appropriate directory.
9. Click Analyze, and watch for a pop-up when it completes. Depending on the number of catchments, delineation can take a long time.
10. After acknowledging the process has completed you may get a second pop-up saying that it was unable to delineate a number of catchments. You may have to delineate these manually.
11. Copy the output file back to your computer and load it into ArcMap, along with the local hydrology, catchments, HUCs, snapped points, and a base map.
12. Perform initial quality control checks (see section below).
13. Compare the delineated catchments to the sites BaseFile to find out which catchments have not been delineated.
14. Manually delineate those catchments that failed automatic delineation or were rejected during quality control checks in an edit session.
    a. Recalculate the New_Lat and New_Long in case there were any changes to the point locations.
15. Once all catchments have been reviewed and delineated, project the shapefile into NAD_1983_California_Teale_Albers using the project tool in the ArcToolbox

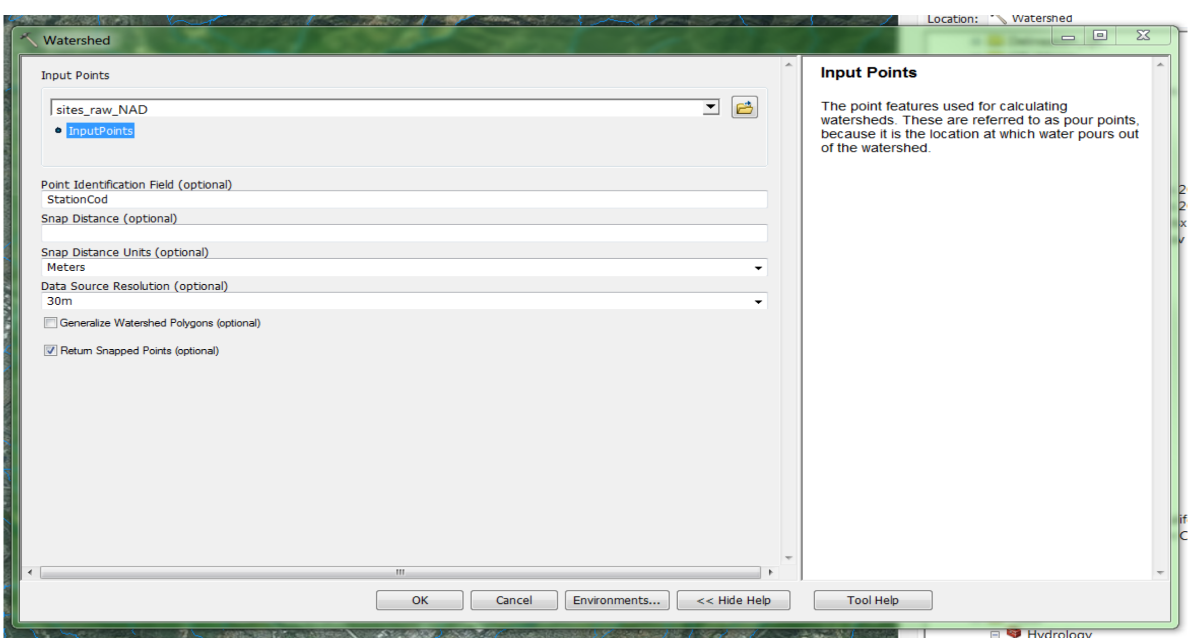


a.

16. Export the shapefile and name them appropriately to make a permanent layer. Name this shapefile XXX_WS (where XXX is the project name).

ArcGIS Hydrology

1. Input parameters:
   - Site coordinate shapefile in "Input Points"
   - Select the unique identification field (StationCod)
   - Leave snapping distance empty
   - Standard data resolution is 30m
   - Do not check the "Generalize Watershed Polygons" option
   - Check "Return Snapped Points"



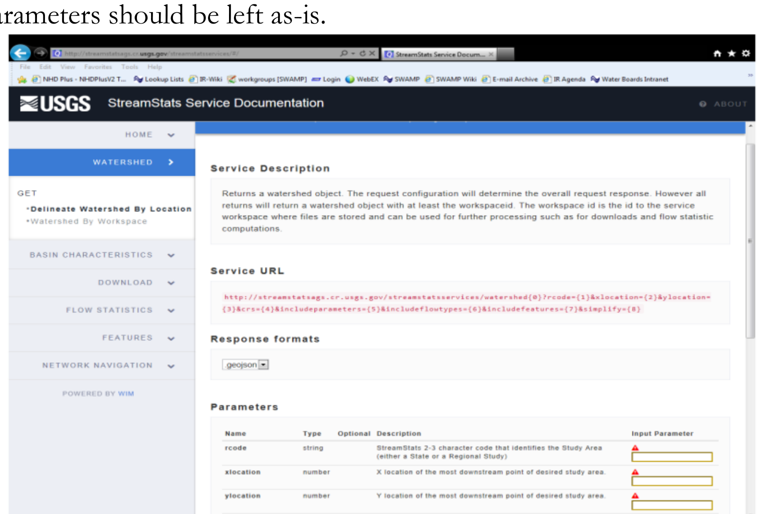
2. Export “Output Snapped Points” and “Output Watershed” outputs
3. Check the delineations for any errors/failures
   - If a snapped point is too far from the actual point, subset original point data that resulted in erroneous snapping and re-run the Watershed tool with an adjusted “Snap Distance” less than the distance to the nearest convergence. Default snapping distance is the specified resolution multiplied by 5.
   - If a delineation fails, manually snap the original points to the NHD line and re-run the Watershed tool.

4. Create new coordinate and ID fields
   - Calculate New_Lat in points file: Add new field (data type Double). Select field header - Calculate Geometry – Property = Y coordinate of point, Coordinate system = NAD 1983 (2011) California (Teale) Albers (Meters), Unit = Decimal Degrees
   - Calculate New_Long in points file: Add new field (type = Double). Select field header - Calculate Geometry – Property = X coordinate of point, Coordinate system = NAD 1983 (2011) California (Teale) Albers (Meters), Unit = Decimal Degrees
   - Update Shape_Leng: select field header - Calculate Geometry – Property = Perimeter, Unit = Meters (m)
   - Update Shape_Area: select field header - Calculate Geometry – Property = Area, Unit = Square Meters (sq m)
   - Add StationCod field: Add field – select field header – Field Calculator – Double click PourPtID in the “Fields” section

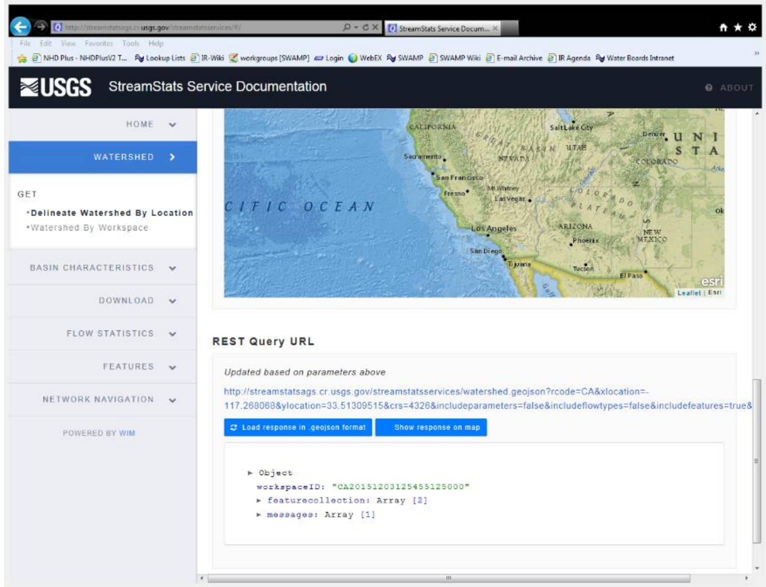
Stream Stats:

1. For sites not near a NHDFlowline, delineation must be done manually. We do not yet have a manual delineating process in place. Depending on the case, edits on the NHD may be required.
   a. Another option is to use USGS Stream Stats (http://streamstatsags.cr.usgs.gov/streamstatsservices/#/) may be used to delineate the watershed.
      i. To delineate a watershed using USGS Stream Stats:
         1. Fill in CA for rcode, longitude for xlocation, and latitude for ylocation. All other parameters should be left as-is.

a.

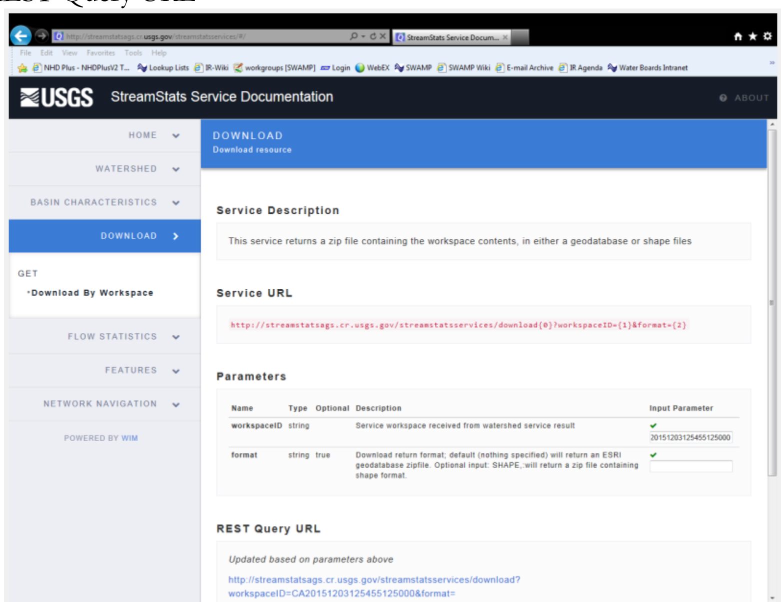


2. Click on "Load response in .geojson format" button. Once done loading, copy the workspaceID



a.

3. Paste unique numbers into workspaceID field and then click on hyperlink below REST Query URL



a.

4. Copy the output file back to your computer and load it into ArcMap, along with the local hydrology, catchments, HUCs, snapped points, and a base map.
5. Perform initial quality control checks (see Quality Control Checks for Catchment Delineations section).

Quality Control Checks for Catchment Delineations

Helpful GIS files to support QC:

- NHD Plus stream network. You may want to hide pipelines, but keep canals visible with a distinct color. **Note: the NHD (1:24K) network is often needed to resolve discrepancies between NHD+ hydrology and DEM based hydrology. If there is a conflict, the 1:24K version is usually much more accurate.**
- Elevation files, shaded relief maps or topographic maps.

1. In general, it is best to examine each catchment individually. Highlighting (or selecting) each catchment, one at a time, makes many problems obvious.
2. Look for gross irregularities, such as:
    - Holes **Fix holes** (this is a pretty rare problem). Do this by removing the polygon vertices that create the hole.
    - Small nonsensical polygons that clearly don't correspond to a drainage network. These tend to occur when the coordinates plot off of a stream line and/or when the stream is in a flat area with little or no relief.
3. Does the watershed have a "lollipop" or "frying pan" shape? This problem is most common when the site is located in a flat area with few topographic features. Unless this shape is supported by the local topography, **flag the site for further review.** Use the catchments from the corresponding regions NHDPlus as guide to fixing "lollipops". Select and merge catchments to delineated catchments where necessary to fill out catchment, or manually correct.
4. For sites close to confluences (within ~300 m), make sure that the "correct" catchment was delineated. The only way to verify this may be to check the original site name or description, or to check with the original field crew that sampled the site.
5. Follow the perimeter of the delineation around the entire watershed. Note the following potential errors:
    - Does the delineation cross any ponds, reservoirs, or lakes? If so, does the topography support inclusion in/exclusion from the watershed? **Fix, or flag for further review**.
    - Do any NHD Plus flowlines cross the watershed border? If so, does the topography support inclusion in /exclusion from the watershed? Flowlines that represent pipelines, canals or aqueducts (or any situation where the flowline does not receive water from the immediate landscape) should be ignored. If necessary, check site with imagery from Google Earth. **Fix, or flag for further review**.
    - Most errors are small, and will have negligible influence on PHAB scores or other predictors. As a rule of thumb, errors can be ignored if they would modify the total area of the catchment <5%, and do not alter the type of landuse inside the delineation.
    - Watch for "divots" in the catchment. If the hydrology does not connect to the rest of the hydrologic network it will not be included in the catchment by the delineator even if they clearly feed into the catchment. Select the NHD Plus catchment and merge it into the delineated catchments in this case.

CALCULATING PREDICTOR METRICS

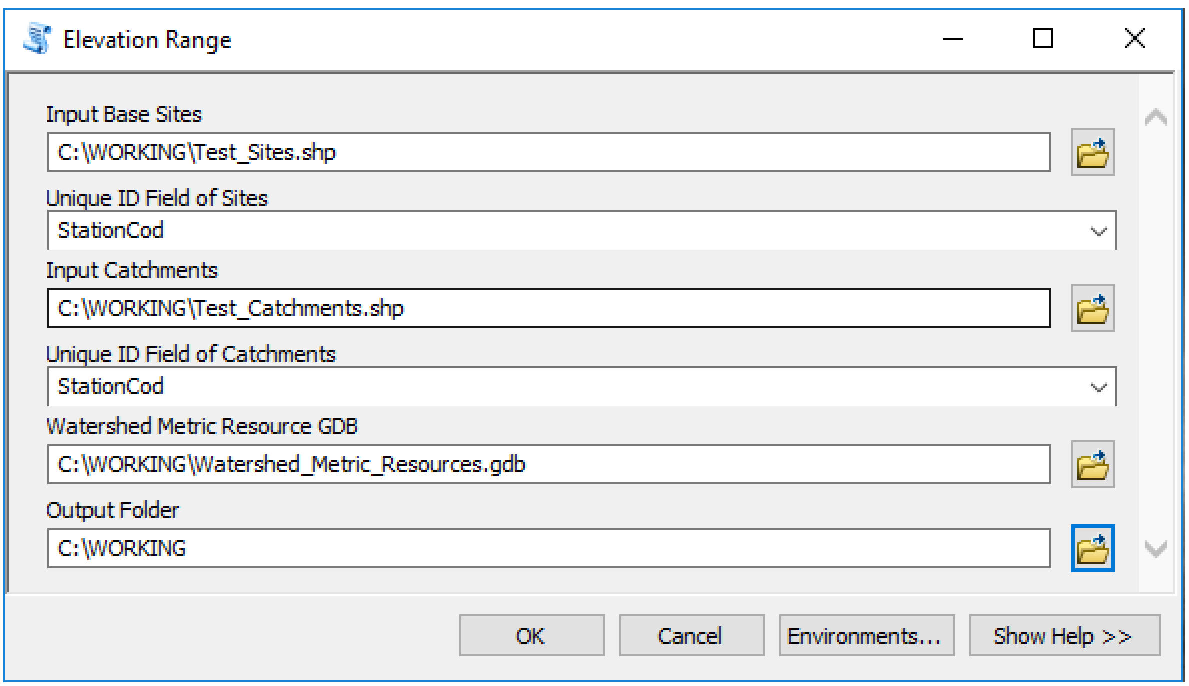
Elevation Metrics

The following describes how to process site elevation, watershed maximum elevation and the elevation change between them using the Elevation Range Python Tool in ArcGIS 10.2.2 and above. This tool requires the Spatial Analyst Extension to run.

Elevation Range Processing Tool

1. Add "Watershed Metric Toolbox" to your ArcToolbox
    a. right-clicking ArcToolbox
    b. select "Add Toolbox"
    c. browse to Watershed Metric Toolbox on your computer and click Open

2. Navigate to the "CSCI Toolset" and double-click the "Elevation Range" script to open its dialog box.



Input Base Sites: Add the site points to this input.

Unique ID Field of Sites: Choose the field that contains the unique id for each input site point. The "Input Catchments" input much have the same set of unique id values. In this example both "test_site_points.shp" and "test_catchments.shp" both contain the field "StationCod".

Input Catchments: Add the catchments polygons to this input. The "StationCod" field must correspond with the Input Base Sites for the tool to run properly.

Unique ID Field of Catchments: Choose the field that contains the unique id for each input catchment polygon. The "Input Base Sites" input much have the same set of unique id values. In this example both "test_site_points.shp" and "test_catchments.shp" both contain the field "StationCod".

Watershed Metric Resource GDB: This is the GDB input with the DEM layers.

Output Folder: Choose the location you wish the final results shapefiles to be saved. Intermediate files will also be saved here during processing but will be deleted upon completion.

3. Click "OK" and the tool will run. When it completes you should see two new shapefiles named "Catchments_Elevation_Ranges.shp" and "Sites_Elevation.shp".

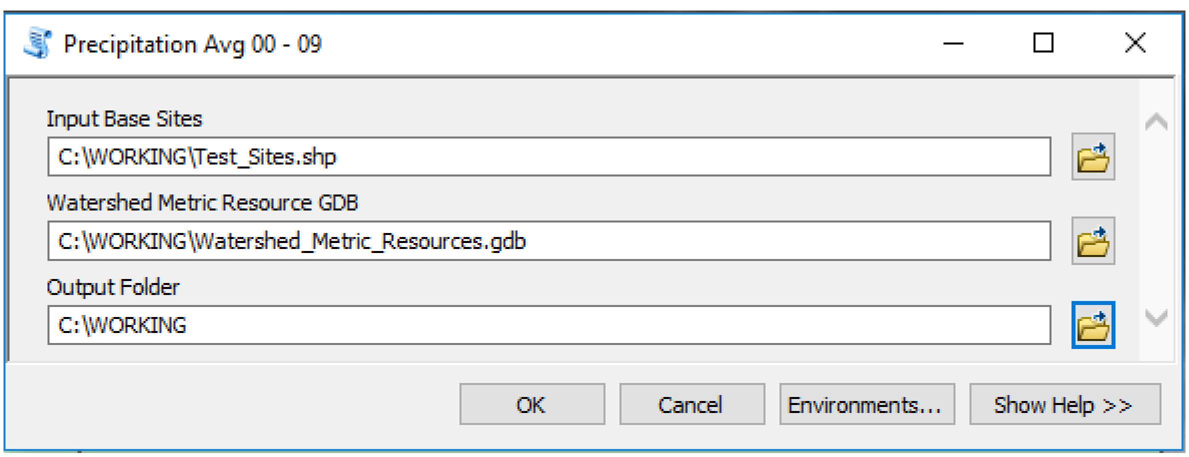
4. Add the new shapefile, "Catchment Elevation Ranges.shp" to ArcMap and open the attribute table.
    a. The follow new fields are added.
        i. SITE_ELEV – Elevation at the sample site in meters.
        ii. MAX_ELEV – Maximum elevation of the watershed in meters.
        iii. ELEV_RANGE – Range between the site and maximum elevations in watershed.
        iv. AREA_SQKM – Watershed area in square kilometers.

Average Precipitation

The following describes how to derive the average precipitation at a giving test site using the Precipitation Avg Python Tool in ArcGIS 10.2.2 and above. This tool requires the Spatial Analyst Extension to run.

<u>Precipitation Processing Tool</u>

1. Navigate to the "CSCI Toolset" and double-click the "Precipitation Avg 00 - 09" script to open its dialog box.



Input Base Sites: Add the site points to this input.

Watershed Metric Resource GDB: This is the geodatabase to holds all of the resource data for the PHAB tools.

Output Folder: Choose the location you wish the final results shapefile to be saved.

2. Click "OK" and the tool will run. When it completes you should see a new shapefile named "PPTAvg_wgs84.shp". Add the new shapefile to ArcMap and open the attribute table. You will see that a new field "PPT_00_09" has been added. It contains the average precipitation from 2000 to 2009 for each site.

   The precipitation units are millimeters multiplied by 100.

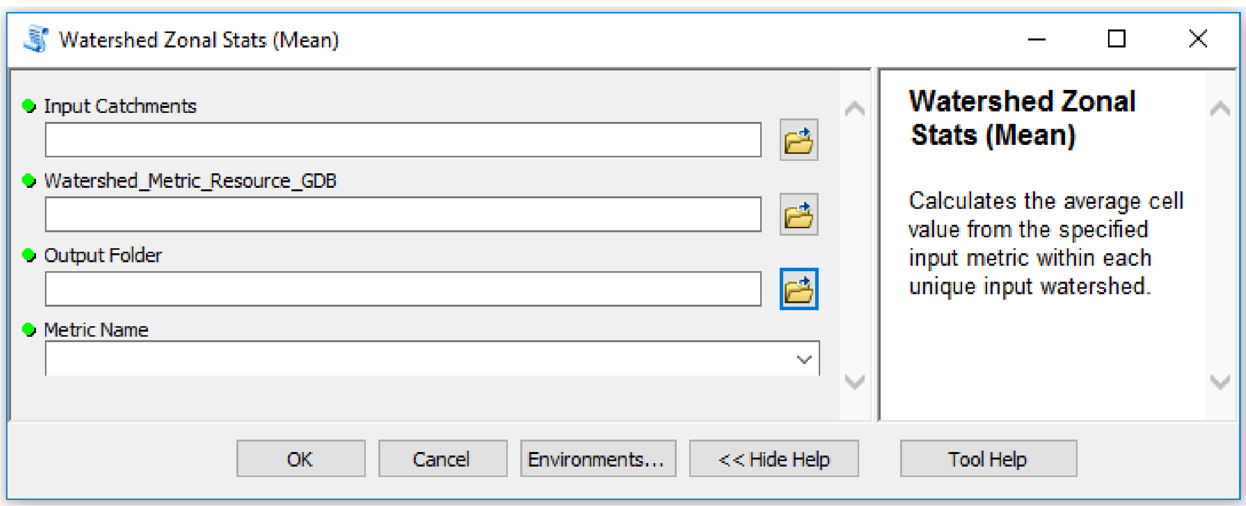
Notes: Currently to script output field assumes the standard 2000 to 2009 time frame but it can easily be modified to output a new field name for any time frame if new data is acquired.

Zonal Statistics (e.g., geology metrics)

The following section describes how to process average values of any input raster within a watershed using the Watershed Zonal Stats (Mean) Python Tool in ArcGIS 10.2.2. This tool requires the Spatial Analyst extension to run.

Watershed Zonal Statistics Processing Tool

1. Navigate to the "CSCI Toolset" and double-click the "Watershed Zonal Stats (Mean)" script to open its dialog box.



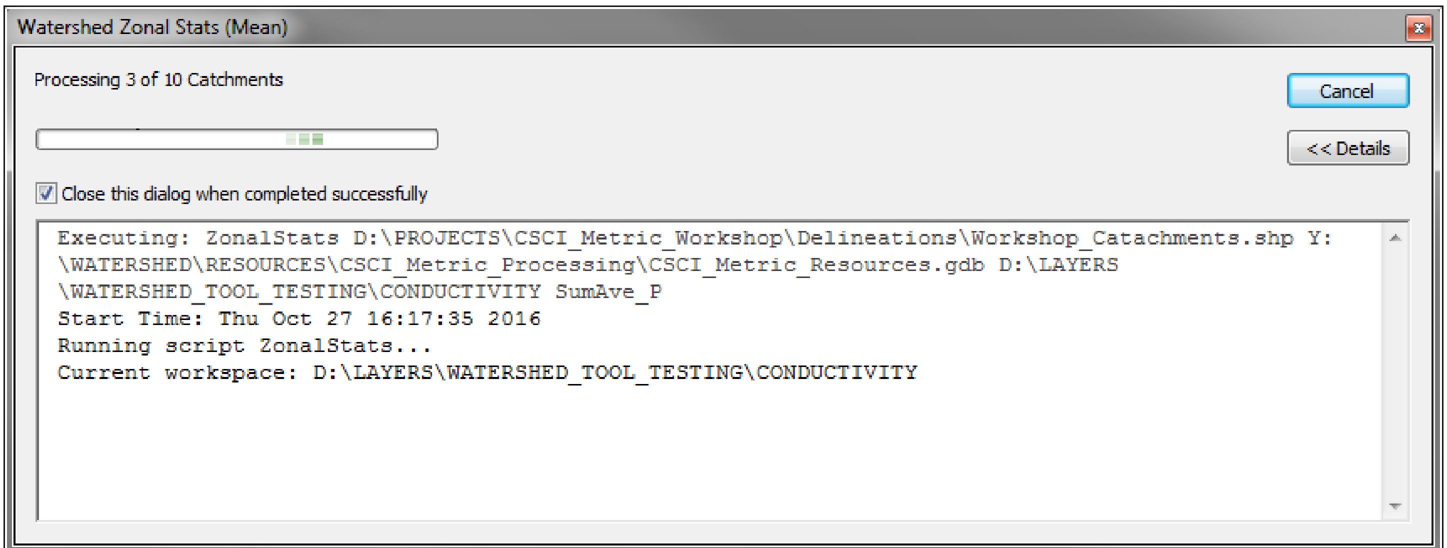
Input Catchments: Add the catchments polygons to this input. They must be projected in California NAD83 Teale Albers

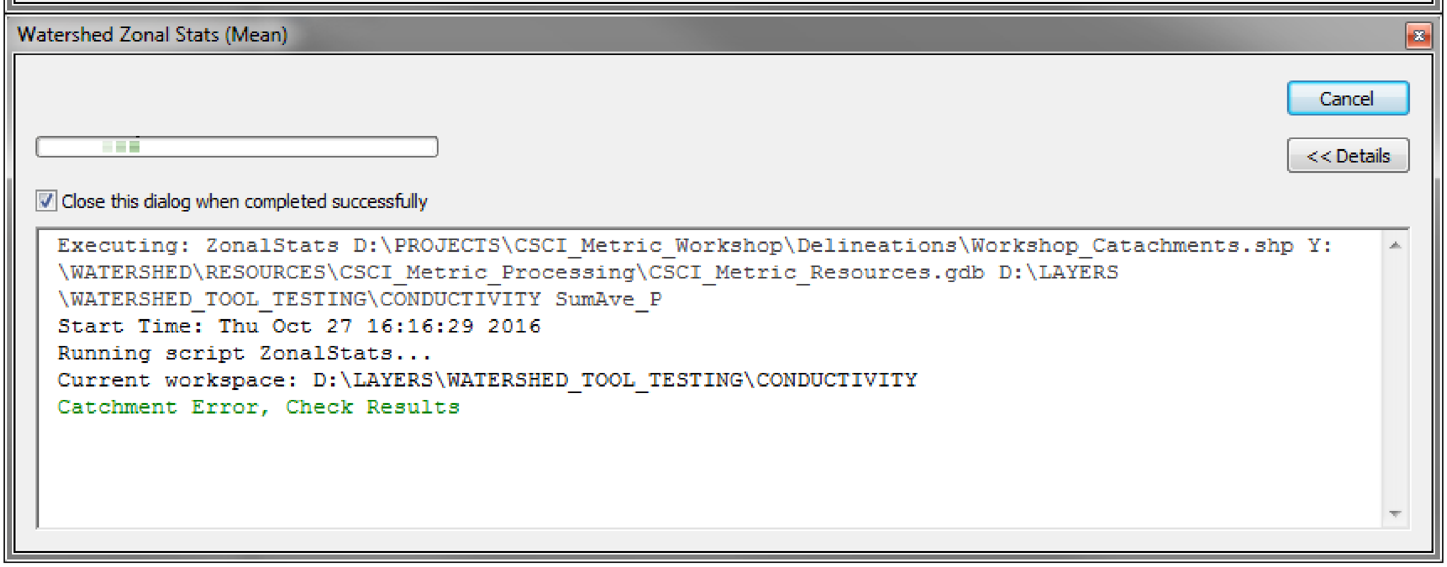
Watershed Metric Resources GDB: This is the geodatabase to holds all of the resource data for the PHAB tools.

Output Folder: Choose the location you wish the final results shapefiles to be saved. Intermediate files will also be saved here during processing but will be deleted upon completion.

Metric Name: From the drop down menu, choose the metric you wish to calculate. This in turn selects the correct input raster from the Watershed Metric Resources GDB and will become the name of the output field in the resultant output dataset. There are many options. For PHAB calculations you need to calculate “MEANP_WS”, “MINP_WS”, “KFCT_AVE”.

Click “OK” and the tool will run. As the tool runs, you will see a progress on how many catchments have been processed out of the total number in your input. In some cases a catchment will not overlap with the input raster, or is too small compared with the input raster cell size. In these cases, the message “Catchment Error, Check Results” is displayed.



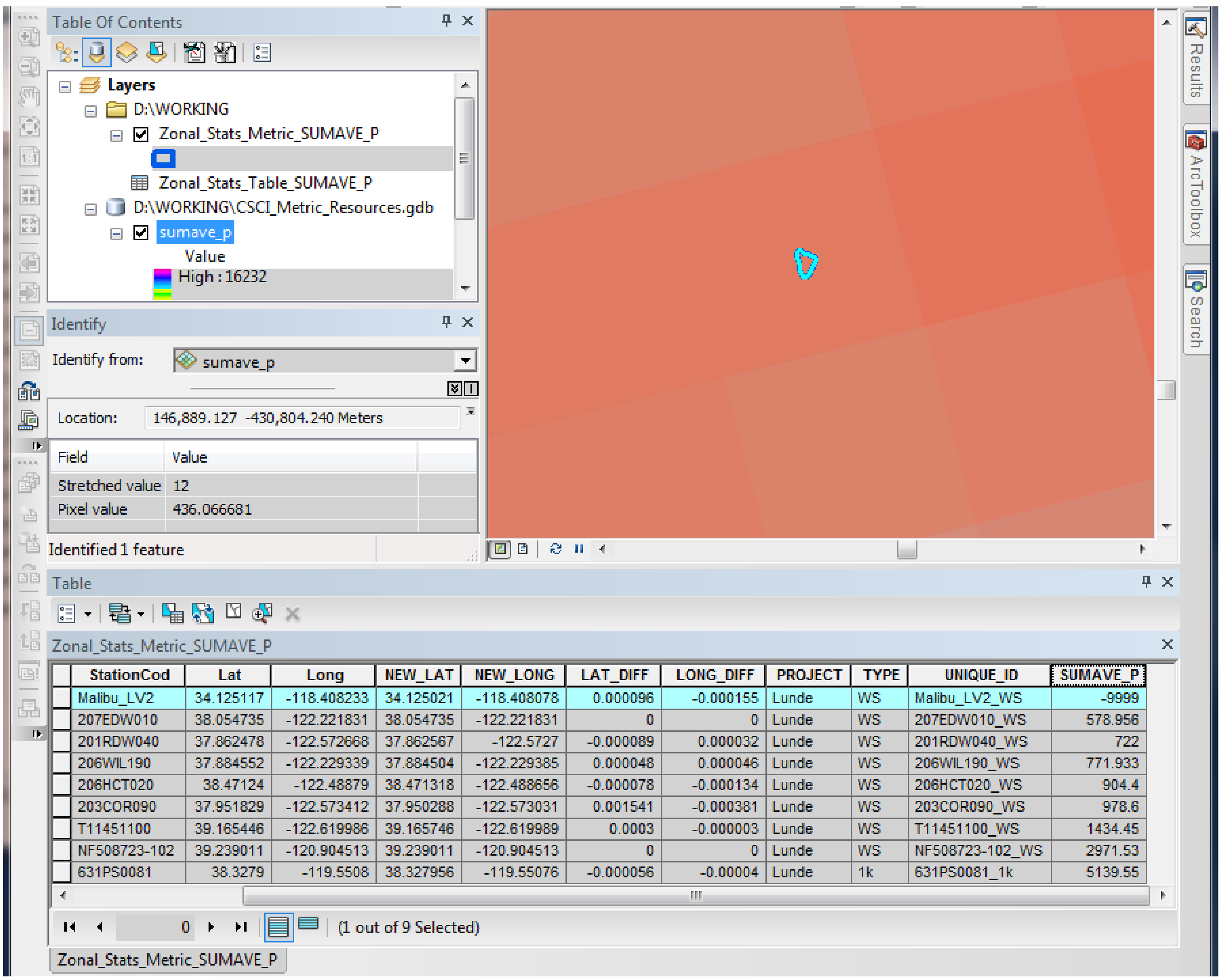


In this example, the catchment “Malibu_LV2” did not process in the model properly. It will be assigned the value -9999. Each catchment given the value -9999 will need to be manually reviewed to determine proper action. This process will be explained in more detail in steps 4 through 7.

3. When the tool completes you will have a shapefile named “Zonal_Stats_Metric_<*Statistic Field Name*>.shp. The Statistics Field Name chosen is used to name your output file. Add the file to ArcMap and open the attribute table to review.

4. Sort ascending on the Statistic field. Check for any values of -9999. If none are present then your data is complete. In the example below the catchment “Malibu_LV2” was assigned the value -9999.

5. Add the Input Stats Raster to ArcMap and zoom to “Malibu_LV2”. In the example below we can see that the catchment was too small for the Zonal Statistics operation to run properly against the “sumave_p” raster.



6. Use the Identify tool to determine the value of the raster at “Malibu_LV2”; in this case it’s 436.066681. Use the field calculator to replace -9999 with the raster cell value.

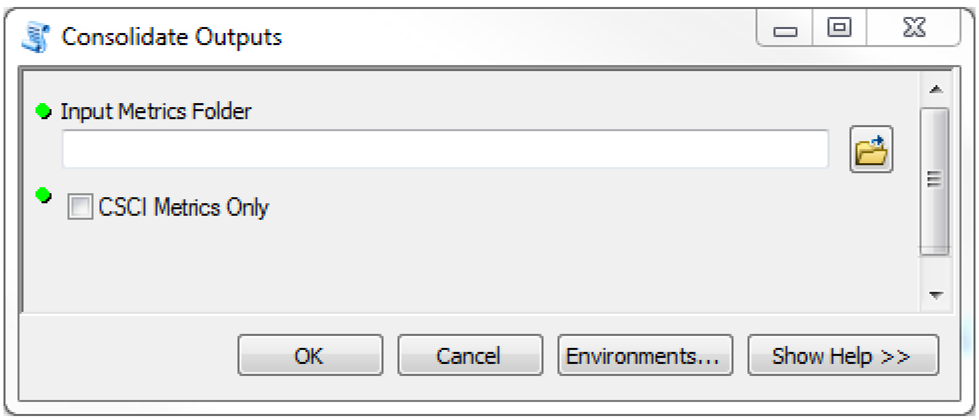
7. Repeat this process for any catchment that has a value of -9999. If the catchment does not overlap with the raster, assign the value from the raster cell closest to the catchment.

METRIC CONSOLIDATION AND DATA EXPORT

The following describes how to merge all results from other tools into a single CSV file ready for input into the R model using the Consolidate Outputs Python Tool in ArcGIS 10.2.2 and above.

Consolidate Metrics Processing Tool

1. Navigate to the "Watershed Metric Toolbox" and double-click the "Consolidate Outputs" script to open its dialog box.



   Input Metrics Folder: Add the folder location were you have saved all of your PHAB outputs. They must all be contained within the same folder.

   CSCI Metrics Only: Make sure the box is unchecked when consolidating PHAB metrics.

2. Click "OK" and the tool will run. When the tool completes a CSV file named "All_Metrics_Consolidated.csv" is generated.

The Consolidate Outputs tool does not check that all PHAB Metrics are present in the chosen folder location. If any are missing they will not be included in the output .csv file. The tool may also not consolide the files correctly if the shapefile "Catchment Elevation Ranges.shp" does not include the columns "StationCod", "New_Lat", and "New_Long". These fields may have to be added manually.

Section 2: Instructions for Calculating IPI Scores with R

The PHAB package contains materials to calculate an index of physical integrity (IPI) for California streams. This index is estimated using site-specific data and available metrics of physical habitat to describe overall integrity.

This tutorial assumes that the user is familiar with basic operations in the R programming language, such as data import, export, and manipulation. Although not required, we recommend using an integrated development environment for R, such as R-studio, which can be downloaded at http://www.rstudio.com. New users are encouraged to pursue training opportunities, such as those hosted by local R user groups. A list of such groups may be found here: http://blog.revolutionanalytics.com/local-r-groups.html. R training material developed by SCCWRP can also be found online: https://sccwrp.github.io/SCCWRP_R_training/

INSTALLATION

Install the package as follows:

```
install.packages('devtools')
library(devtools)
install_github('SCCWRP/PHAB')
library(PHAB)
```

BASIC USAGE

The core function is `IPI` that requires `stations` and `phab` data as input.

```
IPI(stations, phab)
```

```
##   StationCode SampleDate       PHAB_SampleID  IPI IPI_percentile
## 1   105PS0107  9/14/2009 105PS0107_9/14/2009 1.16           0.90
## 2   205PS0157  6/19/2012 205PS0157_6/19/2012 1.04           0.62
## 3   305PS0057  6/16/2009 305PS0057_6/16/2009 0.79           0.04
## 4   504PS0147  6/23/2008 504PS0147_6/23/2008 0.78           0.03
## 5   632PS0007  7/23/2008 632PS0007_7/23/2008 1.10           0.79
## 6   901PS0057  5/14/2008 901PS0057_5/14/2008 1.08           0.74
##   Ev_FlowHab Ev_FlowHab_score H_AqHab H_AqHab_pred H_AqHab_score H_SubNat
## 1       0.85             0.89    1.59         1.11          1.00     1.57
## 2       0.96             1.00    1.42         1.35          0.80     1.41
## 3       0.50             0.51    1.32         1.42          0.70     0.49
## 4       0.28             0.28    1.24         1.30          0.72     0.98
## 5       0.90             0.95    1.51         1.41          0.82     1.80
## 6       0.70             0.73    1.52         1.38          0.84     1.80
##   H_SubNat_score PCT_SAFN PCT_RC PCT_SAFN_pred PCT_SAFN_score XCMG
## 1           0.83        6      0         24.60           1.00   99
## 2           0.74       51      2         22.06           0.40  131
## 3           0.26       83      0         29.51           0.12  152
## 4           0.52        1      0         34.38           1.00   55
## 5           0.95       14      0         13.49           0.79  126
```

```
## 6            0.95       40      3         34.46          0.69  122
##   XCMG_pred XCMG_score IPI_qa Ev_FlowHab_qa H_AqHab_qa H_SubNat_qa
## 1     93.64       0.78   1.00             1          1           1
## 2    104.72       0.90   1.00             1          1           1
## 3    106.05       1.00   1.00             1          1           1
## 4     95.41       0.51   1.00             1          1           1
## 5    123.10       0.76   1.00             1          1           1
## 6    102.16       0.86   0.95             1          1           1
##   PCT_SAFN_qa XCMG_qa
## 1           1    1.00
## 2           1    1.00
## 3           1    1.00
## 4           1    1.00
## 5           1    1.00
## 6           1    0.95
```

DETAILED USAGE

The PHAB package can be installed from the R console with just a few lines of code. The current version of the package can be found on SCCWRP's GitHub page here and can be installed using the devtools package. The devtools package must be installed first before the PHAB package can be installed. Start by installing and loading devtools:

```
install.packages('devtools')
library(devtools)
```

Now the `install_github()` function from devtools can be used to install PHAB from GitHub. The package can be loaded after installation.

```
install_github('SCCWRP/PHAB')
library(PHAB)
```

The installation process may take a few seconds. Both the devtools and PHAB packages depend on other R packages, all of which are installed together. If an error is encountered during installation, an informative message is usually printed in the R console. This information can help troubleshoot the problem, such as identifying which dependent packages may need to be installed separately. Please contact SCCWRP staff if additional errors are encountered.

After the package is successfully installed, you will be able to view the help file for the PHAB core function:

```
?IPI
```

Preparing Input Data

The IPI is estimated using station and PHAB metric data as input. Station GIS data can be obtained using the GIS instructions that accompany this document and PHAB metric data can be obtained from the state of California SWAMP reporting module. The SWAMP reporting module calculates a suite of PHAB metrics for data stored in the SWAMP database; options to calculate metrics for data outside the SWAMP database (e.g., CEDEN) are currently in development. Sample data are provided with the PHAB package to demonstrate the required format and are loaded automatically once the package is installed and loaded (i.e., just type the name of the sample data in the R console to view). You can view the `stations` and `phab` example data from the R console:

```
head(stations)
```

```
##   StationCode MAX_ELEV  AREA_SQKM ELEV_RANGE MEANP_WS  New_Long SITE_ELEV
## 1   105PS0107     2587 2002.90397    2050.96 882.4179 -123.0173    536.04
## 2   205PS0157     1111  600.95071    1074.00 604.9312 -121.8352     37.00
## 3   305PS0057     1152 1261.70747    1107.36 544.3272 -121.5114     44.64
## 4   504PS0147     2144 1954.62572    2059.52 781.2332 -122.2178     84.48
## 5   632PS0007     3348   98.66497    1240.06 962.0929 -119.5937   2107.94
## 6   901PS0057     1734   92.31964    1671.43 495.7878 -117.6696     62.57
##    KFCT_AVE  New_Lat MINP_WS PPT_00_09
## 1 0.1776000 41.71375 10.9294   54996.9
## 2 0.2927333 37.30141  0.7059   40023.9
## 3 0.2840000 36.95052  0.6623   42432.2
## 4 0.2181000 39.77572  2.7024   53740.0
## 5 0.1217000 38.53335 17.9569   67846.9
## 6 0.2854000 33.52814  0.5268   29440.0
```

```
head(phab)
```

```
##   StationCode SampleDate      Variable Result Count_Calc
## 1   105PS0107  9/14/2009  W1_HALL_EMAP   0.00         11
## 2   105PS0107  9/14/2009 W1_HALL_SWAMP   0.00         14
## 3   105PS0107  9/14/2009      PCT_CPOM  24.00        105
## 4   105PS0107  9/14/2009      Ev_AqHab   0.77          8
## 5   105PS0107  9/14/2009    Ev_FlowHab   0.85          4
## 6   105PS0107  9/14/2009     Ev_SubNat   0.75          8
```

The `stations` data include site-specific information derived from geospatial analysis. These data are in wide format where one row corresponds to data for one site. The following fields are required:

- `StationCode` unique station identifier
- `MAX_ELEV` maximum elevation in the watershed in meters
- `AREA_SQKM` watershed area in square kilometers
- `ELEV_RANGE` elevation range of the watershed in meters
- `MEANP_WS` mean phosphorus geology from the watershed
- `New_Long` site longitude, decimal degrees
- `SITE_ELEV` site elevation
- `KFCT_AVE` average soil erodibility factor
- `New_Lat` site latitude, decimal degrees
- `MINP_WS` minimum phosphorus geology from the watershed
- `PPT_00_09` average precipitation (2000 to 2009) at the sample point, in hundredths of millimeters

The phab data include calculated physical habitat metrics that are compiled along with the `stations` data to get the IPI score. These data are in long format where multiple rows correspond to physical habitat metric values for a single site. The following fields are required:

- `StationCode` unique station identifier
- `SampleDate` date of the sample
- `Variable` name of the PHAB metric
- `Result` resulting metric value
- `Count_Calc` number of unique observations that were used to estimate the value in `Result`

Values in the `Variable` column of the phab data indicate which PHAB metric was measured that corresponds to values in the `Result` column. The required PHAB metrics that should be provided for every unique sampling event specified by `StationCode` and `SampleDate` are as follows:

- `XSLOPE` mean slope of reach
- `XBKF_W` mean bankfull width
- `H_AqHab` Shannon diversity of aquatic habitat types
- `PCT_SAFN` percent sand and fine (<2 mm) substrate particles
- `XCMG` riparian cover sum of three layers
- `Ev_FlowHab` evenness of flow habitat types
- `H_SubNat` Shannon diversity of natural substrate types
- `XC` mean upper canopy trees and saplings
- `PCT_POOL` percent pools in reach
- `XFC_ALG` mean filamentous algae cover
- `PCT_RC` percent concrete/asphalt

Each PHAB metric serves a specific purpose in the calculation and reporting of the IPI. Some metrics (i.e., `H_AqHab`, `PCT_SAFN`, `XCMG`, `Ev_FlowHab`, and `H_SubNat`) are used to assess habitat condition. Other metrics (i.e., `XSLOPE`, `XBKF_W`, and `PCT_RC`) are used as predictors or score modifiers for components of the IPI. Finally, some metrics (i.e., the `XC`, `PCT_POOL`, and `XFC_ALG`) are required because they are used for quality assurance checks.

All required fields for the `stations` and phab data are case-sensitive and must be spelled correctly. The order of the fields does not matter. All `StationCode` values must be shared between the datasets. As described below, the `IPI()` function automatically checks the format of the input data prior to estimating scores.

Detailed Metric Descriptions

Five of the required PHAB metrics in the input data are used directly for scoring the IPI, whereas the remainder serve a supporting role as predictors or modifiers for different parts of the complete index. Understanding what each of five core metrics describe about stream condition and how they vary with disturbance is critical for interpreting the index. Below is a detailed description of each metric (excerpted from the tech memo, click for more detail).

Shannon diversity of natural instream cover. `H_AqHab` measures the relative quantity and variety of natural structures in the stream, such as cobble, large and small boulders, fallen trees, logs and branches, and undercut banks available as refugia, or as sites for feeding or spawning and nursery functions of aquatic macrofauna. A wide variety and/or abundance of submerged structures in the stream provides macroinvertebrates and fish with a large number of niches, thus increasing habitat diversity. When variety and abundance of cover decreases (e.g., due to hydromodification, increased sedimentation, or active stream clearing), habitat structure becomes monotonous, diversity decreases, and the potential for recovery following disturbance decreases. Snags and submerged logs—especially old logs that have remained in-place for several years–are among the most productive habitat structure for macroinvertebrate colonization and fish refugia in low-gradient streams.

Percent sand and fine substrate. `PCT_SAFN` measures the amount of small-grained sediment particles (i.e., <2 mm) that have accumulated in the stream bottom as a result of deposition. Deposition may result from soil disturbance in the catchment, landslides, and bank erosion. Sediment deposition may cause the formation of islands or point bars, filling of runs and pools, and embeddedness of gravel, cobble, and boulders and snags, with larger substrate particles covered or sunken into the silt, sand, or mud of the stream bottom. As habitat provided by cobbles or woody debris becomes embedded, and as interstitial spaces become inundated by sand or silt, the surface area available to macroinvertebrates and fish is decreased. High levels of sediment deposition are symptoms of an unstable and continually changing environment that becomes unsuitable for many organisms. Although human activity may deplete sands and fines (e.g., by upstream dam operations), and this depletion may harm aquatic life, the IPI treats only increases in this metric as a negative impact on habitat quality, although a *post-hoc* correction was made whereby the metric percent concrete (`PCT_RC`) is added to `PCT_SAFN` before scoring.

Shannon diversity of natural substrate types. `H_SubNat` measures the diversity of natural substrate types, assessing how well multiple size classes (e.g., gravel, cobble and boulder particles) are represented. In a stream with high habitat quality for benthic macroinvertebrates, layers of cobble and gravel provide diversity of niche space. Occasional patches of fine sediment, root mats and bedrock also provide important habitat for burrowers or clingers, but do not dominate the streambed. Lack of substrate diversity, e.g., where >75% of the channel bottom is dominated by one particle size or hard-pan, or with highly compacted particles with no interstitial space, represents poor physical conditions. Riffles and runs with a diversity of particle sizes often provide the most stable habitat in many small, high-gradient streams.

Evenness of flow habitat types. `Ev_FlowHab` measures the evenness of riffles, pools, and other flow microhabitat types. Optimal physical conditions include a relatively even mix of velocity/depth regimes, with regular alternation between riffles (fast-shallow), runs (fast-deep), glides (slow-shallow) and pools (slow-deep). Poor conditions occur when a single microhabitat dominates (usually glides, with pools and riffles absent). A stream that has a uniform flow regime will typically support far fewer types of organisms than a stream that has a variety of alternating flow regimes. Riffles in particular are a source of high-quality habitat and diverse fauna, and their regular occurrence along the length of a stream greatly enhances the diversity of the stream community. Pools are essential for many fish and amphibians.

Riparian vegetation cover, sum of three layers. `XCMG` measures the amount of vegetative protection afforded to the stream bank and the near-stream portion of the riparian zone. The root systems of plants growing on stream banks help hold soil in place, thereby reducing the amount of erosion likely to occur. The vegetative zone also serves as a buffer to pollutants entering a stream from runoff and provides shading and habitat and nutrient input into the stream. Banks that have full, multi-layered, natural plant growth are better for fish and macroinvertebrates than are banks without vegetative protection or those shored up with concrete or riprap. Vegetative removal and reduced riparian zones occur when roads, parking lots, fields, lawns, bare soil, riprap, or buildings are near the stream bank. Residential developments, urban centers, golf courses, and high grazing pressure from livestock are the common causes of anthropogenic degradation of the riparian zone. Even in undeveloped areas, upstream hydromodification and invasion by non-native species can reduce the cover and quality of riparian zone vegetation.

Using the `IPI()` Function

The IPI score for a site is estimated from the station and PHAB data. The score is estimated automatically by the `IPI()` function in the package following several steps. First, reference expectations for a site are estimated for predictive metrics using the station data. Then, observed data values are compared to reference expectations for predictive metrics and the differences between observed and predicted (i.e., metric residuals) are used for scoring. For metrics that are not predicted, raw metric values are used for scoring. Metric scores are based on the upper and lower percentiles of either metric residuals or raw metric values observed at reference and high-activity sites. The metric scores are then summed and standardized (i.e., divided) by the mean sum of scores at reference sites to obtain the final IPI score.

The `IPI()` function can be used on station and PHAB data that are correctly formatted as shown above. The `stations` and `phab` example data are in the correct format and are loaded automatically with the PHAB package. These data are used here to demonstrate use of the `IPI()` function.

```
IPI(stations, phab)
```

```
##   StationCode SampleDate       PHAB_SampleID  IPI IPI_percentile
## 1   105PS0107  9/14/2009 105PS0107_9/14/2009 1.16           0.90
## 2   205PS0157  6/19/2012 205PS0157_6/19/2012 1.04           0.62
## 3   305PS0057  6/16/2009 305PS0057_6/16/2009 0.79           0.04
## 4   504PS0147  6/23/2008 504PS0147_6/23/2008 0.78           0.03
## 5   632PS0007  7/23/2008 632PS0007_7/23/2008 1.10           0.79
## 6   901PS0057  5/14/2008 901PS0057_5/14/2008 1.08           0.74
##   Ev_FlowHab Ev_FlowHab_score H_AqHab H_AqHab_pred H_AqHab_score H_SubNat
## 1       0.85             0.89    1.59         1.11          1.00     1.57
## 2       0.96             1.00    1.42         1.35          0.80     1.41
## 3       0.50             0.51    1.32         1.42          0.70     0.49
## 4       0.28             0.28    1.24         1.30          0.72     0.98
## 5       0.90             0.95    1.51         1.41          0.82     1.80
## 6       0.70             0.73    1.52         1.38          0.84     1.80
##   H_SubNat_score PCT_SAFN PCT_RC PCT_SAFN_pred PCT_SAFN_score XCMG
## 1           0.83        6      0         24.60           1.00   99
## 2           0.74       51      2         22.06           0.40  131
## 3           0.26       83      0         29.51           0.12  152
## 4           0.52        1      0         34.38           1.00   55
## 5           0.95       14      0         13.49           0.79  126
## 6           0.95       40      3         34.46           0.69  122
##   XCMG_pred XCMG_score IPI_qa Ev_FlowHab_qa H_AqHab_qa H_SubNat_qa
## 1     93.64       0.78   1.00             1          1           1
## 2    104.72       0.90   1.00             1          1           1
## 3    106.05       1.00   1.00             1          1           1
## 4     95.41       0.51   1.00             1          1           1
## 5    123.10       0.76   1.00             1          1           1
## 6    102.16       0.86   0.95             1          1           1
##   PCT_SAFN_qa XCMG_qa
## 1           1    1.00
## 2           1    1.00
## 3           1    1.00
```

```
## 4           1    1.00
## 5           1    1.00
## 6           1    0.95
```

A data frame of IPI scores estimated at each site on each unique sample date is returned. The output data are in wide format with one row for each sample date at a site. Detailed information for each output column is as follows:

- `StationCode` unique station identifier
- `SampleDate` date of the site visit
- `PHAB_SampleID` unique identifier of the sampling event. Typically, the station code and sample data are sufficient to determine unique sampling events.
- `IPI` score for the index of physical integrity
- `IPI_percentile` the percentile of the IPI score relative to scores at reference sites
- `Ev_FlowHab` evenness of flow habitat types, from the raw PHAB metric
- `Ev_FlowHab_score` IPI score for evenness of flow habitat types
- `H_AqHab` Shannon diversity of natural instream cover types, from the raw PHAB metric
- `H_AqHab_pred` predicted Shannon diversity of natural instream cover types
- `H_AqHab_score` scored Shannon diversity of natural instream cover types
- `H_SubNat` Shannon Diversity of natural substrate types, from the raw PHAB metric
- `H_SubNat_score` scored Shannon diversity of natural substrate types
- `PCT_SAFN` percent sand and fine substrate, from the raw PHAB metric
- `PCT_RC` percent concrete/asphalt, from the raw PHAB metric and is combined with `PCT_SAFN` to provide an overall estimate of substrate with poor suitability for macrofauna
- `PCT_SAFN_pred` predicted percent sand and fine substrate
- `PCT_SAFN_score` scored percent sand and fine substrate, includes `PCT_RC`
- `XCMG` riparian cover as sum of three layers, from the raw PHAB metric
- `XCMG_pred` predicted riparian cover as sum of three layers
- `XCMG_score` scored riparian cover as sum of three layers
- `IPI_qa` quality assurance for the IPI score
- `Ev_FlowHab_qa` quality assurance for evenness of flow habitat types
- `H_AqHab_qa` quality assurance for Shannon diversity of aquatic habitat types
- `H_SubNat_qa` quality assurance for Shannon diversity of natural substrate types
- `PCT_SAFN_qa` quality assurance for percent sand and fine substrate
- `XCMG_qa` quality assurance for riparian cover as sum of three layers

Metrics are included in the output as observed PHAB metrics, predicted metrics (where applicable), and scored metrics. Observed PHAB metrics are returned as-is from the input data. Some PHAB metrics include a predicted column that shows the modelled metric value based on the environmental setting at a site. Scored PHAB metrics are obtained following the description above.

The last five columns include quality assurance information for the IPI score and select metrics. QA values less than one indicate less quality assurance, usually resulting from metric values being calculated from fewer measurements from a sample than specified by field protocols. IPI_qa (the overall QA value for the IPI) is based on the lowest score among all metrics. At this time, there is no criterion for flagging an IPI score based on QA measurements; analysts are advised to use their personal judgment when evaluating IPI scores with low QA measurements.

Interpreting IPI Scores

The IPI was calibrated during its development so that the mean score of reference sites is 1; IPI scores near 1 represent locations with conditions similar to reference sites. Scores that approach 0 indicate great departure from reference condition and degradation of physical condition. Scores > 1 can be interpreted to indicate greater physical complexity than predicted for a site given its natural environmental setting. All metric scores are weighted equally to determine the overall IPI score. For observed and scored PHAB metrics, all are expected to decrease under degraded physical conditions, except `PCT_SAFN` which is expected to increase in response to degradation.

Calibration Data

An additional data file is available within the PHAB package that shows calibration data for scoring the IPI metrics. This file is called `refcal` and includes observed and predicted scores at reference and high-activity (or "stressed") sites for the five PHAB metrics. Metrics are scored based on deviation from the 5th and 95th percentile of scores at reference or calibration sites. The `refcal` dataset includes observations at these sites that were used to identify percentile cutoffs for estimating metric scores (the dataset is loaded automatically with the PHAB package, just type the name in the console to view).

```
head(refcal)
```

```
##     Variable StationCode      SampleID2   SiteSet Result Predicted
## 1 Ev_FlowHab   000CAT228 000CAT22840400    RefCal   0.77 0.6498149
## 2 Ev_FlowHab   101WE1111 101WE111137474 StressCal   0.94 0.6535123
## 3 Ev_FlowHab   103CDCHHR 103CDCHHR40435    RefCal   0.63 0.7295132
## 4 Ev_FlowHab   103WER026 103WER02637831    RefCal   0.75 0.6590854
## 5 Ev_FlowHab   103WER029 103WER02937832    RefCal   0.94 0.7125525
## 6 Ev_FlowHab   105BVCAGC 105BVCAGC40442    RefCal   0.84 0.6973526
```

- `Variable` name of the PHAB metric
- `StationCode` unique station identifier
- `SampleIDs` unique identifier of the sampling event
- `SiteSet` indicating if a site was reference or stressed
- `Result` resulting metric value
- `Predicted` predicted metric value

Error Checks for Input Data

The `IPI()` function will evaluate both the `stations` and `phab` input datasets for correct format before estimating IPI scores. IPI scores will not be calculated if any errors are encountered. The following checks are made:

- No duplicate station codes in `stations`. That is, input data have one row per station.
- All station codes in `stations` are in `phab`, and vice-versa
- All required fields are present in `stations` and `phab` (see above)

- All required PHAB metrics are present in the `variable` field of `phab` for each station and sample date (see above). An exception is made for `XC`, `PCT_POOL`, and `XFC_ALG`, which are not necessary for calculating the IPI but are used for optional quality assurance checks.
- No duplicate results for PHAB variables at each station and sample date. That is, one row per station, date, and phab metric.
- All input variables for `stations` and `phab` are non-negative, excluding elevation variables in `stations` which may be negative if below sea level (i.e., some locations in southeast California). Moreover, the variables `XBKF_W` and `Ev_FlowHab` in `phab` must also be greater than zero.

The `IPI()` function will print informative messages to the R console if any of these errors are encountered. It is the responsibility of the analyst to correct any errors in the raw data before proceeding.